

Relationship between Origin and Genetic Diversity in Chinese Soybean Germplasm

Yiwu Chen and Randall L. Nelson*

ABSTRACT

The soybean [*Glycine max* (L.) Merr.] was domesticated in China. Information about the amount and distribution of genetic diversity in China is critical to effective soybean germplasm management. Information is currently available from only a few provinces in China. The objectives of this research are to estimate the genetic variation within and among four geographically diverse provinces (Zhejiang, Sichuan, Gansu, and Hebei) in China and to determine the relationship between geographical origin and genetic diversity. Ten primitive cultivars from each province were characterized by random amplified polymorphic DNA (RAPD) fragments produced from 31 selected decamer primers. CNS, an important U.S. ancestral line, was also included as a control. Genetic variation was estimated by AMOVA analysis with 125 polymorphic RAPD fragments. Genetic distances were calculated by means of Jaccard's coefficient and expressed as dissimilarity coefficients. Unweighted paired group method using arithmetic averages (UPGMA), Ward's minimum-variance method, VARCLUS, and multidimensional scaling (MDS) were applied to define the genetic relationships. AMOVA identified significant genetic differences between all pairs of provinces except between Zhejiang and Sichuan. The greatest difference was observed between Hebei and Zhejiang. There was disagreement among the clustering methods, but each procedure identified clusters of accessions that originated from the same province. Based on data from all clustering procedures, six major clusters containing a total of 32 accessions were defined with each cluster dominated by accessions from a single province. These data provide additional evidence that primitive cultivars of China were generally genetically isolated in relatively small geographical areas.

SOYBEAN originated in China and has been cultivated for more than 3000 yr. The long history of cultivation in different environments has contributed to the evolution of many genetically distinct soybean types in China. There are over 22 600 accessions in the Chinese soybean germplasm collection that represent every province except Tibet and Qinghai (Chang and Sun, 1991; Chang et al., 1996). On the basis of the data from the catalogs of Chinese soybean collection, 93% of these accessions are primitive cultivars, which tend to be highly diverse and provide ideal materials for genetic diversity studies. Developing a better understanding of the diversity of these important genetic resources is critical for effective

management and utilization of soybean germplasm. Knowing the range and distribution of the genetic diversity of a species can affect acquisition priorities for a germplasm collection and sampling strategies for germplasm screening research or for establishing a core collection. It also allows breeders to better understand the evolutionary relationships among accessions and to develop strategies to integrate useful diversity into their breeding programs (Bretting and Widrechner, 1995).

Several diversity studies in soybean have been conducted using morphological characters, pedigree information, and biochemical variation (Nelson et al., 1987, 1988; Gizlice et al., 1994; Sneller, 1994; Bernard et al., 1998). Although morphological and agronomic characters are useful in evaluating genetic diversity, collecting such data can be laborious and the phenotypic values are often strongly influenced by the environment. Biochemical variants such as isozymes and electrophoresis patterns of storage proteins are less affected by the environment but have limited variation. DNA markers are an attractive alternative. They are nearly unlimited in numbers, presumably selectively neutral, and can be organized into linkage maps (Thormann and Osborn, 1992).

There are several reports of using molecular markers for evaluation of primitive soybean germplasm. RAPD markers have been shown to be a simple and effective means to evaluate variability in crops. RAPDs are well suited for diversity studies because they are technically simple, nonradioactive, relatively inexpensive, and require small amounts of DNA. On the basis of principal component analysis of RAPD fragment data from 35 diverse soybean lines, Thompson and Nelson (1998) established a core set of RAPD primers with high polymorphism in soybean. These 35 core RAPD primers have been used in other studies for genetic diversity analysis in soybean (Brown-Guedira et al., 2000; Li et al., 2001; Li and Nelson, 2001, 2002). Thompson et al. (1998) evaluated 18 U.S. soybean ancestors and 17 selected Chinese accessions from the USDA Soybean Germplasm Collection. The genetic relationships defined by RAPD data generally corresponded to known pedigrees, origins, and maturity groups. Li and Nelson (2001, 2002) examined the genetics relationships among eight to 10 accessions from each of Henan, Hebei, Shaanxi, Ningxia, Heilongjiang, Shandong, Jiangsu, and Shanxi provinces in China. On the basis of the results of several clustering analyses, they found that the groups formed generally reflected the geographical origin of the accessions. Brown-Guedira et al. (2000) did not find an association between origin and RAPD markers among soybean lines of more modern origin. It is highly likely that these genotypes have been dispersed by human intervention from the areas of actual origin.

Abbreviations: AMOVA, analysis of molecular variance; MDS, multidimensional scaling; MG, maturity group; PCR, polymerase chain reaction; PI, plant introduction; RAPD, random amplified polymorphic DNA; UPGMA, unweighted pair group method using arithmetic average.

Yiwu Chen, Dep. of Crop Sciences, 1101 W. Peabody Dr., University of Illinois, Urbana, IL 61801; Randall L. Nelson, USDA-Agricultural Research Service, Soybean/Maize Germplasm, Pathology, and Genetics Research Unit, Dep. of Crop Sciences, 1101 W. Peabody Dr., University of Illinois, Urbana, IL 61801. Mention of a trademark, proprietary product, or vendor does not constitute a guarantee or warranty of the product by the USDA or the University of Illinois and does not imply its approval to the exclusion of other products or vendors that may also be suitable. Received 5 Feb. 2004. *Corresponding author (rlnelson@uiuc.edu).

Published in Crop Sci. 45:1645–1652 (2005).
Plant Genetic Resources
doi:10.2135/cropsci2004.0071
© Crop Science Society of America
677 S. Segoe Rd., Madison, WI 53711 USA

Patterns of RAPD markers also have been shown to be associated with geographical origin in barley (*Hordeum vulgare* L.) (Fernandez et al., 2002), chickpea (*Cicer arietinum* L.) (Iruela et al., 2002), common bean (*Phaseolus vulgaris* L.) (Beebe et al., 2000), groundnut (*Vigna subterranea* L.) (Amadou et al., 2001), wild emmer wheat (*Triticum dicoccoides* (Koern. Ex Asch. & Graebner) Aaronson) (Fahima et al., 1999), durum wheat (*Triticum turgidum* L. var. *durum*) (Spagnoletti Zeuli and Qualset, 1993), and mango (*Mangifera indica* L.) (Lopez-Valenzuela et al., 1997). Yee et al. (1999) did not find a relationship between RAPD marker patterns and geographical origin for azuki (*Vigna angularis* Willd.). They speculated that two factors may have influenced these results: the genepool of cultivated azuki is quite restricted and the putative origin data used in their analysis may not have been correct. RAPD marker variation was not related to geographical origin in the cultivated races of sorghum [*Sorghum bicolor* (L.) Moench] (Menkir et al., 1997), which may be the result of high levels of gene flow

among the regions (Aldrich and Doebley, 1992). The objectives of this research were to estimate the genetic variation of primitive soybean cultivars within and among four geographically diverse provinces (Zhejiang, Sichuan, Gansu, and Hebei) in China and to determine the relationship between geographical origin and genetic diversity by using RAPD markers.

MATERIALS AND METHODS

Ten randomly selected primitive cultivars (cultivars that pre-date scientific breeding) from 10 geographically and uniformly distributed counties in each of Hebei, Gansu, Sichuan, and Zhejiang provinces in China were used in this research (Table 1 and Fig. 1). These provinces were selected to complement the work of Li and Nelson (2001) by adding information from additional provinces and evaluating a different set of accessions from Hebei. They also represent a wide range of latitudes, elevations, and ecological zones within China and the origin of the oldest Chinese agricultural civilization (Wang, 1982a). The soybean germplasm accessions from these provinces account

Table 1. The origin of the 41 soybean genotypes used in this study and genetic associations determined on the basis of the statistical analyses of RAPD fragments.

Code†	Entry	PI number	Origin		Latitude	MG‡	GD§	Groups based on cluster analyses			
			Province	County				VAR¶	AVE#	Ward's††	Cluster‡‡
G10	Tu huang dou	PI 567351B	Gansu	Minqin	38.6° N	III	0.27	1	2	2	1
H3	Hei bai men huang dou	PI 567492	Hebei	Baxian	39.1° N	IV	0.32	1	2	2	1
	CNS	PI 548445	Jiangsu			VII	0.30	1	2	2	1
S1	Gan gu huang-2	PI 587966A	Sichuan	Shizhu	30.0° N	VIII	0.30	1	1	2	1
Z7	Ba yue bai	PI 587894	Zhejiang	Shaoxing	30.0° N	VII	0.25	1	2	2	1
Z9	Shan bai dou	PI 587904	Zhejiang	Tiantai	29.2° N	VII	0.31	1	2	2	1
Z10	Tian geng dou	PI 587934	Zhejiang	Pingyang	27.6° N	X	0.31	1	2	2	1
G5	He se huang dou	PI 567302	Gansu	Gaotai	39.6° N	I	0.28	3	3	1	2
H6	Tu er dun	PI 567504	Hebei	Xinlong	40.5° N	II	0.35	3	3	1	2
H7	Tu er yan	PI 567505	Hebei	Pingquan	41.0° N	II	0.32	3	3	1	2
H5	Huang dou	PI 567501	Hebei	Dongguang	37.9° N	IV	0.30	3	3	1	2
H4	Huang dou	PI 567497	Hebei	Xinle	38.3° N	III	0.27	3	4	3	3
H10	Zhua zi tui huang dou	PI 567513	Hebei	Baoding	38.8° N	III	0.30	3	4	3	3
H1	Er huang dou	PI 567490	Hebei	Gucheng	37.4° N	IV	0.29	3	4	3	3
H9	Xiao li bai dou	PI 567509	Hebei	Chengnan	36.5° N	IV	0.36	3	4	3	3
H8	Xiao huang dou-2	PI 567507E	Hebei	Xingtai	37.0° N	V	0.29	3	4	3	3
G2	Bian huang dou	PI 567294	Gansu	Huachi	36.6° N	IV	0.35	4	3	1	4
G4	Chang man you huang dou	PI 567299A	Gansu	Gangu	34.8° N	V	0.26	4	3	1	4
G3	Chan yao dou	PI 567298	Gansu	Huating	35.4° N	V	0.29	4	3	1	4
G9	Niu mao huang	PI 567345	Gansu	Kanxian	33.4° N	V	0.27	4	3	1	4
G8	Ma huang dou	PI 567343	Gansu	Zhouqu	33.8° N	V	0.27	4	3	1	4
H2	Er huang yang	PI 567491A	Hebei	Yuxian	39.8° N	III	0.28	4	3	1	4
S5	Da li dong dou	PI 587987A	Sichuan	Nanchong	31.0° N	IV	0.35	4	3	1	4
S7	Liu yue huang	PI 587991	Sichuan	Mianyang	31.5° N	IV	0.27	5	2	5	5
S2	Liu yue bao	PI 587967	Sichuan	Yuyang	29.0° N	IV	0.28	5	2	2	5
S4	Zhuang zhuang dou	PI 587983A	Sichuan	Junlian	28.0° N	IV	0.29	5	2	2	5
S9	Da bai dou	PI 588024B	Sichuan	Yanyuan	27.6° N	V	0.27	5	2	2	5
S10	Da huang ke	PI 588027A	Sichuan	Yuxi	28.7° N	V	0.27	6	2	2	6
Z2	Duan jia ai jiao huang	PI 587853	Zhejiang	Jiaxing	30.7° N	VII	0.28	6	2	4	6
Z6	Qing pi dou	PI 587889	Zhejiang	Taishun	27.6° N	VIII	0.24	6	2	2	6
Z4	Ba yue dou	PI 587874	Zhejiang	Kaihua	29.2° N	VIII	0.30	6	2	4	6
Z1	Zei mo xiao	PI 587849	Zhejiang	Changxin	31.0° N	VIII	0.24	6	2	4	6
G1	Bai huang dou	PI 567291	Gansu	Baiyin	36.6° N	IV	0.28	2	1	5	7
G7	Hua lai dou	PI 567318	Gansu	Ningxian	35.5° N	IV	0.26	2	3	4	7
G6	Hei pi gu huang dou	PI 567311A	Gansu	Lintao	35.5° N	IV	0.27	2	5	5	7
S3	Xiao huang dou	PI 587977	Sichuan	Zigong	29.5° N	IV	0.28	2	2	5	7
S6	Shuang hua huang jiao dou	PI 587989A	Sichuan	Dayi	30.8° N	IV	0.32	2	3	4	7
S8	Bai huang dou	PI 587993B	Sichuan	Guangyuan	32.5° N	VII	0.27	2	2	4	7
Z3	Huang pi dou	PI 587870	Zhejiang	Tonglu	29.8° N	VII	0.25	2	5	5	7
Z8	Qing pi dou	PI 587901	Zhejiang	Suichang	28.6° N	VII	0.30	2	2	4	7
Z5	Ba yue huang	PI 587884	Zhejiang	Dinghai	30.0° N	VII	0.26	2	2	4	7

† G = Gansu, H = Hebei, S = Sichuan, and Z = Zhejiang.

‡ MG = U.S. maturity group.

§ GD = Mean genetic distance calculated from Jaccard's coefficient from all of the other accessions.

¶ VAR = Cluster defined by VARCLUS.

AVE = Cluster defined by Unweighted Paired Group Method using Arithmetic Averages (UPGMA).

†† Ward's = Cluster defined by Ward's minimum-variance method.

‡‡ Cluster = Cluster assignment made based on the results of all clustering procedures.

for 20% of entire Chinese germplasm collection as determined on the basis of Chinese soybean germplasm catalogs (Wang, 1982b; Chang and Sun, 1991; Chang et al., 1996). Gansu, in northwest China, is located between 33 and 42° N with elevations ranging from 500 to 2000 m. Some northern parts of this province are desert. Hebei is east of Gansu between 36 and 42° N. There are mountainous areas on the northern and western side of the province with elevations of approximately 1000 m. The remainder of the province is a large plain less than 200 m above sea level. Sichuan, in southwest China, is between 26 and 34° N. The large central area of the province known as Sichuan Valley has an elevation of approximately 200 m. It is surrounded by mountains which reach 1000 m on the east and more than 2000 m on the west. Zhejiang, on the east coast of China, is the most uniform ecological environment among these four provinces. It is located at 27 to 31° N with an average elevation of 200 m. U.S. maturity group was not considered in the selection process but ranged from MG I to V in Gansu, MG II to V in Hebei, MG IV to VIII in Sichuan, and MG VII to X in Zhejiang. CNS, a major U.S. ancestral line that originated from Jiangsu province, was also included as a control since the RAPD banding patterns in CNS have been well documented in previous research (Brown-Guedira et al., 2000; Li et al., 2001; Li and Nelson, 2001, 2002).

First trifoliate leaves were harvested from approximately 10 greenhouse-grown seedlings. Leaf pieces were placed in 1.5-mL microcentrifuge tubes and cooled in liquid nitrogen for 2 min. DNA was extracted by the CTAB (hexadecyltrimethyl ammonium bromide) method of Keim et al. (1988). The DNA concentration of all extracted samples was calculated by spectrophotometer readings at wavelengths of 260/280 and adjusted to a concentration of 10 ng/μL.

Thirty-one decanucleotide primers from Operon Technologies Inc. (Alameda, CA) were chosen for this study. Thompson and Nelson (1998) had shown all of these primers to be highly

polymorphic in diverse soybean germplasm. The protocol of Kresovich et al. (1994) with minor modifications was used for amplification. A reaction mixture of 25 μL with approximately 50 ng of genomic DNA was amplified in an Omnigene thermal cycler (Hybaid Inc., Franklin, MA). Amplified products were separated by 1% (w/v) agarose gels in 1× Tris-acetate buffer, stained with ethidium bromide, and visualized under UV light.

RAPD fragments were scored as either present (1) or absent (0). The Jaccard's coefficient was used to calculate the similarity coefficients between each pair of genotypes for all polymorphic loci using the following formula: $S_{ij} = a/(a + b + c)$, where a is the number of common bands (1, 1); b is the number of bands present in the first accession and absent in the second (1, 0); and c is the number of bands absent in the first accession and present in the second (0, 1). The 0, 0 matches were not considered useful because the lack of a RAPD bands in two genotypes may not be due to a common mutational event. Genetic distances were reported as dissimilarity coefficients, $D_{ij} = 1 - S_{ij}$. The matrix of genetic distances was submitted to two hierarchical procedures, unweighted pair group method using arithmetic averages (UPGMA) and Ward's minimum-variance methods (SAS Institute, 1999) to cluster the entries. Values of the cubic clustering criterion (CCC), pseudo F statistic (PSF), and Hotelling's pseudo T^2 statistic were used to define the optimum cluster numbers as described in the SAS STAT user's manual (SAS Institute, 1999). A nonhierarchical cluster analysis procedure, VARCLUS option of PROC CLUSTER in PC SAS (SAS Institute, 1999), was also applied to the original fragment data to divide the accessions into non-overlapping clusters. The matrix of genetic distances generated from genetic dissimilarity coefficients was subjected to multi-dimensional scaling (MDS) (Shepard, 1974) by the MDS procedure in PC SAS (SAS Institute, 1999). To maintain the scale of 0 and 1 for making interpretation and graphing easier the ABSOLUTE option was used. The criteria are similar to that

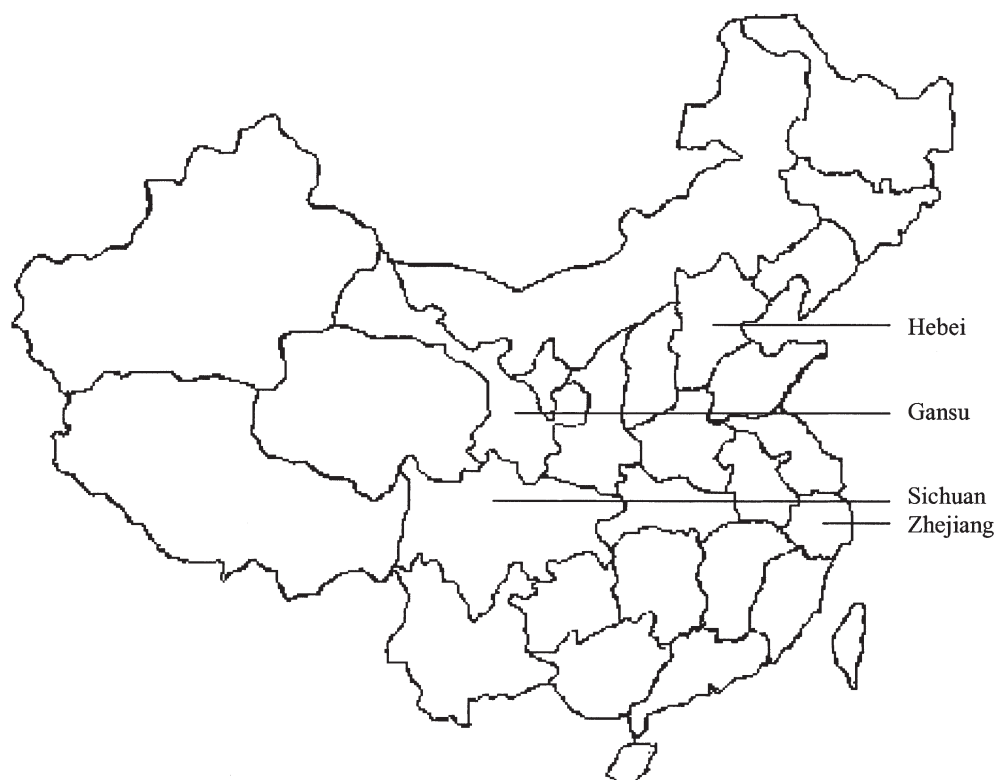


Fig. 1. Location of the four provinces of China from which the 40 selected soybean accessions originated.

described by Gizlice et al. (1996) and Thompson et al. (1998). To evaluate the effectiveness of 2 to 22 dimensions, the goodness-of-fit criterion (R^2) between the original data and the predicted values that were derived from the MDS coordinates was used. The best MDS analysis was considered to be the fewest dimensions that result in $R^2 > 0.95$ with the original genetic distance matrix. Polymorphism information content (PIC) scores (Anderson et al., 1993) were determined by the formula, $PIC_i = 1 - \sum P_{ij}^2$ (Weir, 1990) where P_{ij} is the frequency of the j th marker allele for fragment i . The raw data were also subjected to principal component analysis. This involves a mathematical procedure to transform a set of correlated response variables into a smaller set of uncorrelated variables called principal components.

The components of variance attributable to differences among provincial populations, and among individuals within provinces were estimated from the genetic distance matrix, as specified in the analysis of molecular variance (AMOVA) procedure in ARLEQUIN ver 2.000 (Schneider et al., 2000). This procedure has been used to analyze genetic diversity in a variety of crops including sweet potato [*Ipomoea batatas* (L.) Lam.] (Zhang et al., 2000), sesame (*Sesamum indicum* L.) (Ercan et al., 2004), maize (*Zea mays* L.) (Reif et al., 2003), olive (*Olea europaea* L.) (Belaj et al., 2002), pearl millet [*Pennisetum glaucum* (L.) R. Br.] (Bhattacharjee et al., 2002), as well as soybean [*Glycine max* (L.) Merr.] (Li and Nelson, 2002). A nonparametric permutation procedure with 3000 permutations was used to test the significance of variance components associated with the different possible levels of genetic structure in this study (Excoffier et al., 1992). The pairwise F_{st} values, a value of F statistic analogs computed from AMOVA, were used to compare genetic distances between any two provinces (Schneider et al., 2000). Since soybean is a diploid, self-pollinated crop, all population samples have the same mating pattern. We partitioned the genotypic variance and not the variance of allele frequencies as for codominant markers.

RESULTS AND DISCUSSION

RAPD Marker Diversity

Thirty-one RAPD primers produced a total of 241 fragments of which 125 were polymorphic. The fragments produced with CNS were used as standards to help identify the fragments from the other accessions. All of the scored fragments in CNS were found in the other accessions. An average of 7.7 fragments could be scored unambiguously for each primer with a range of 3 to 15 (Table 2). The RAPD primer OPO-05 had the most polymorphic fragments (9) with approximate molecular weights of 500, 550, 700, 950, 1050, 1300, 1400, 2500 and 2800 bp. The percentage of polymorphism (52%) among these primitive cultivars of diverse origin is higher than reported by Thompson and Nelson (1998) for ancestral lines of U.S. cultivars and selected *G. max* exotic accessions (30%) but lower than Li and Nelson (2001) data for both *G. max* and *G. soja* Siebold & Zucc. (56%). Genetic diversity for a specific locus can be measured by polymorphism information content (PIC) scores. The higher the PIC score, the higher the probability that polymorphism will exist between two accessions at that locus if the locus only has two alleles. Because all RAPD fragments are either present or absent, 0.50 will be the highest PIC score for any fragment. The range of PIC scores for polymorphic fragments was 0.05 to 0.50 with

Table 2. The sequences of 31 primers used to characterize the genetic diversity of 41 *G. max* accessions and the number of fragments produced by each primer.

Primers	Sequence	Total number of fragments	Number of polymorphic fragments
	5'→3'		
OPA-20	AATCGGGCTG	5	3
OPE-01	CCCAAGGTCC	6	4
OPF-04	GGTGATCAGG	6	3
OPG-06	GTGCCTAACC	7	2
OPH-02	TCGGACGTGA	7	3
OPH-12	ACGCGCATGT	8	6
OPH-15	AATGGCGCAG	4	1
OPK-01	CATTGAGGCC	8	4
OPK-03	CCAGCTTAGG	13	5
OPK-10	GTGCAACGTG	8	5
OPL-09	TGCGAGAGTC	9	3
OPL-18	ACCACCCACC	12	6
OPN-03	GGTACTCCCC	3	3
OPN-08	ACCTCAGCTC	6	0
OPN-09	TGCCGGCTTG	10	6
OPN-18	GGTGAGGTCA	6	4
OPO-01	GGCACGTAAG	3	3
OPO-04	AAGTCCGCTC	5	2
OPO-05	CCCAGTCACT	15	9
OPO-14	AGCATGGCTC	10	5
OPO-19	GGTGCACGTT	8	5
OPP-07	GTCCATGCCA	8	2
OPR-07	ACTGGCTTGA	10	5
OPR-10	CCATTCCCCA	8	4
OPR-12	ACAGGTGCGT	11	7
OPR-13	GGACGACAAG	6	3
OPS-01	CTACTGCGCT	10	7
OPS-03	CAGAGGTCCC	9	3
OPS-05	TTTGGGGCCT	7	4
OPS-11	AGTCGGGTGG	8	5
OPS-14	AAAGGGGTCC	5	3
Total		241	125

the mean of 0.31. This average PIC score is in general agreement with the result of Thompson and Nelson (1998) and Li and Nelson (2001). The pattern of divergence among the provinces was primarily attributable to differences in fragment frequencies rather than the unique fragments that were found for each province (Table 3). Although the accessions in this study represent only a small portion of the available germplasm from each province, the unique fragments indicate the genetic distinctness of each province.

AMOVA to Partition Genetic Variance among the Populations

On the basis of the AMOVA, the variation among individuals within the four provinces accounted for 90% of the total (Table 4). The variation among provinces accounted for only 10% of the total but was highly significant ($P < 0.0001$). This result is very similar to the conclusions of Li and Nelson (2001) and other genetic diversity studies (Mellish et al., 2002; vom Brocke et al., 2003). In Li and Nelson's study, the variation among provinces was 11% of total, and the variation among the three countries (China, S. Korea, and Japan) only accounted for 12% of the variation. Provincial pairwise comparisons based on the values of F_{st} can be interpreted as standardized interpopulation distances between regional groups. The F statistic (F_{st} value) is used to estimate the correlation of genes of different individuals in the same population and is a measure of genetic dif-

Table 3. RAPD fragments unique to a single Chinese province.

Fragment	PIC (polymorphism information content) scores			
	Hebei	Gansu	Zhejiang	Sichuan
OPA-20 ₁₅₀₀ [†]	0.48	0	0	0
OPG-06 ₈₀₀	0.32	0	0	0
OPL-09 ₁₅₀₀	0.18	0	0	0
OPR-10 ₂₁₀₀	0.18	0	0	0
OPR-12 ₂₃₀₀	0.18	0	0	0
OPS-05 ₅₀₀	0.18	0	0	0
OPS-11 ₄₅₀	0	0.32	0	0
OPF-04 ₂₀₀₀	0	0.18	0	0
OPK-03 ₁₅₀₀	0	0.18	0	0
OPK-03 ₈₇₀	0	0.18	0	0
OPK-03 ₃₅₀	0	0.18	0	0
OPN-03 ₈₅₀	0	0	0.35	0
OPG-06 ₁₇₀₀	0	0	0.32	0
OPH-02 ₄₅₀	0	0	0.32	0
OPK-10 ₁₆₀₀	0	0	0.18	0
OPR-12 ₁₉₀₀	0	0	0.18	0
OPH-02 ₃₀₀	0	0	0	0.18
OPL-18 ₂₁₀₀	0	0	0	0.18
OPL-18 ₁₈₀₀	0	0	0	0.18
OPL-18 ₅₀₀	0	0	0	0.18
OPN-08 ₂₀₀₀	0	0	0	0.42
OPN-09 ₁₆₅₀	0	0	0	0.18
OPN-18 ₄₀₀	0	0	0	0.32
OPO-19 ₁₇₀₀	0	0	0	0.18

[†] Primer designation and approximate molecular weight of specific fragment.

ferentiation over subpopulations. When F_{st} equals 0, the subpopulations are identical in all allele frequencies; when F_{st} equals 1, they are fixed for different alleles. This statistic is very similar to coancestry coefficients (Table 5). The provincial pairwise F_{st} values ranged from 0.02 between the Sichuan and Zhejiang groups to 0.18 between the Hebei and Zhejiang groups (Table 5). All of the pairwise comparisons between provinces are significantly different from zero except the comparison between Sichuan and Zhejiang. Although Sichuan and Zhejiang are geographically separated, both are located at similar latitudes. The genetic distances between the accessions of the northern and southern provinces are all significantly different, indicating a possible relationship between latitude and genetic diversity. Li and Nelson (2001) conducted research on genetic diversity among soybean accessions from China, Japan, and South Korea using RAPD markers and concluded that there was no clear relationship between latitude and genetic diversity among accessions from these three countries. In their study, 110 accessions originated from similar latitudes between 33 to 39° N in the three countries and 10 accessions from Heilongjiang province of China between 44 to 53° N.

Cluster Analyses

To define the optimum number of clusters, we examined the CCC, PSF, and PST² statistics from the output of PROC CLUSTER. The CCC and PSF values indicate that there may be eight clusters whereas the value of PST² indicates six clusters to be the most likely fit for this dataset. Six to eight clusters should be reasonable for these accessions. Accessions that were placed in the same cluster by two or more procedures were assigned to a consensus cluster (Table 1). In general, the clustering procedures produced similar results but the output from some procedures was different. For example, Ward's minimum-variance, UPGMA, and VARCLUS procedures all agreed on the relationships defined in clusters 2, 3, and 4, but clusters 2 and 3 were combined in the VARCLUS procedure, and clusters 2 and 4 were combined in the UPGMA and Ward's procedure.

Cluster 1 contains three accessions from Zhejiang, the latest maturing accession from Sichuan, the U.S. ancestral line CNS, and two much earlier accessions from Gansu and Hebei. CNS was originally selected from an introduction from Jiangsu, which is adjacent to Zhejiang. PI 567351B comes from the northern part of Gansu where high temperatures and low humidity are common. PI 567492 from Hebei comes from the southern plain of the province and has a relatively large seed size (26 g 100 seeds⁻¹). There is no obvious explanation for the inclusion of these lines in this cluster. All procedures agreed on the relationships defined by the assigned cluster except PI 587966A, which was identified as an outlier with UPGMA. In cluster 2, there are three accessions from Hebei including the two earliest lines from that province and the earliest line from Gansu. The two MG II accessions from Hebei originated in a mountainous northern region that is more similar to the environment of Gansu.

Clusters 3 and 5 each contains only accessions from a single province and have a relatively narrow range of maturity. The members of cluster 3 are from the southern plain of Hebei and the members of cluster 5 are from the central valley in Sichuan.

Cluster 4 primarily contains accessions very similar in maturity from the southern part of Gansu. Two lines, PI 567491A from Hebei and PI 587987A from Sichuan, are exceptions to that description. PI 567491A originated from the northern mountainous area of Hebei. On the basis of origin and maturity, PI 587987A seems like a better fit in cluster 5 but none of the procedures grouped PI 587987A with the lines in cluster 5.

Table 4. Analysis of molecular variance results for the analysis of 40 soybean accessions from Hebei, Gansu, Sichuan, and Zhejiang provinces in China.

Source of variation	df	Sum of squares	Expected mean squares [†]	Variance	Percentage of variation	P value
Among provinces	$P - 1 = 3$	145.8	$n\alpha_a^2 + \alpha_b^2$	$\alpha_a^2 = 2.6$	10.1	< 0.0001
Individuals within provinces	$N - P = 36$	832.2	α_b^2	$\alpha_b^2 = 23.1$	89.9	< 0.0001
Total	$N - 1 = 39$	983.0	α_c^2	$\alpha_c^2 = 25.7$	100.0	

[†] α_a^2 is the covariance component of differences among populations; α_b^2 is the covariance component of differences among individuals within populations; n is defined by $n = \frac{N - \sum_p N_p^2}{P - 1}$. In this case $P = 4$, $N = 40$; therefore, $n = 10$.

Table 5. Pairwise comparisons among provinces based on genetic distances between populations and F_{ST} values for 40 soybean accessions from four provinces in China.

	Population pairwise F_{ST} value				Genetic distances based on coancestry coefficients			
	Hebei	Gansu	Sichuan	Zhejiang	Hebei	Gansu	Sichuan	Zhejiang
Hebei	0.000				0.000			
Gansu	0.093	0.000			0.098	0.000		
Sichuan	0.130	0.064	0.000		0.140	0.067	0.000	
Zhejiang	0.178	0.089	0.020NS	0.000	0.200	0.093	0.020NS	0.000

NS, Nonsignificant at 0.05 probability level. All other values are significant at the 0.05 probability level.

† F_{ST} values are measures of genetic differentiation calculated from analysis of molecular variation (AMOVA).

There are four lines from Zhejiang and one line from Sichuan in cluster 6. PI 588027A from Sichuan, MG V, is substantially different in maturity than the lines (PI 587879, PI 587853, PI 587874, and PI 587889) from Zhejiang in MGs VII and VIII. PI 588027A also originated from a much higher elevation than the Sichuan lines in cluster 4 and 5 (PI 587967, PI 587983A, PI 587987A, PI 587991, and PI 588024B).

The remaining nine accessions from Gansu, Sichuan, and Zhejiang were placed in a single cluster by VARCLUS but were not grouped consistently by the other two procedures. On the basis of our criterion, we could have assigned PI 587993B, PI 587901, and PI 587884 to a separate cluster, or we could have placed them into cluster 6. Because of the ambiguity of the data for these accessions, we chose to assign them and the other remaining accessions to cluster 7. From these data, we cannot conclusively determine the relationships among these nine accessions or how they relate to the other major clusters.

There are many similarities in the groupings defined by each of the clustering procedures used in this research. However, if the results from only a single procedure would have been reported the genetic groupings would have been quite different. Using multiple procedures may be an effective way of identifying the most robust clusters.

Because *G. max* is a domesticated species, determining actual genetic origin is often uncertain; however, among most of the accessions in this research there is an association between origin and genetic similarity. Clusters 1 through 6 were each dominated by accessions from a single province. These data provide additional evidence that the primitive cultivars of China were generally genetically isolated in relatively small geographical areas.

Multidimensional Scaling

Twenty-two dimensions were sufficient to retain the information in the original matrix of genetic dissimilarity based on Jaccard's coefficient ($R^2 = 0.99$). Stress, the measure of the amount by which the MDS generated distance mismatched the original dissimilarities, was 2.0% when 22 dimensions were applied in the solution. This level was considered to be a good fit according to the guidelines of Johnson and Wichern (1992). The first two MDS dimensions accounted for 47% of the total variation, whereas only 23% of the total variation was contributed by the first two principle components (data not shown). A plot of the first two MDS dimensions was generated to compare with the results of hierarchical

and nonhierarchical cluster analysis (Fig. 2). The positioning of members of cluster 1 to cluster 6 in the two-dimensional MDS plot was consistent with the assigned clusters (Table 1). The remaining nine accessions are generally located in the upper quadrants of the MDS plot. From the MDS plot it seems that S7 (PI 587991) in cluster 5 is genetically closer to the members of cluster 7 and S6 (PI 587989A) could be a member of cluster 4. The results from Ward's minimum-variance method were in agreement with those classifications (Table 1). From these data and the data from AMOVA and cluster analysis we can conclude that there is a strong relationship between geographical origin and genetic diversity among these primitive cultivars from China.

Implications for Soybean Germplasm Collection

To facilitate the management and use of genetic resources, the concept of a "core collection" was proposed first by Frankel (1984) and later by Brown (1989a) to provide efficient access to the whole collection. A core germplasm collection is a small sample of the accessions in a collection that represents the maximum genetic diversity with minimum repetitiveness. As the number of accessions in a soybean germplasm collection increases, initial evaluation of numerous, highly similar accessions not only wastes resources but also reduces the possibility of identifying the truly unique and valuable accessions. Construction of core collections will be advantageous for plant breeders to initially focus on fewer germplasm accessions. Under sampling theory of selectively neutral alleles, Brown (1989b) recommended a core collection size of about 10% of the entire collection and a number of random sampling strategies of establishing a core collection were proposed. Several studies have shown that phenotypic divergence is related to geographical origin of accessions (Spagnoletti Zeuli and Qualset, 1993; Schoen and Brown, 1995; Lopez-Valenzuela et al., 1997; Fahima et al., 1999; Beebe et al., 2000; Amadou et al., 2001; Li and Nelson, 2001, 2002; Fernandez et al., 2002; Iruela et al., 2002). Thus ecogeographical information can be applied to stratify phenotypic diversity of the entire collection.

On the basis of geographical, morphological, cytological, and isozyme information, Brown et al. (1987) established a core collection of perennial *Glycine*. Extensive databases of geographical, morphological, and agronomic traits on soybean for several major germplasm collections are available to establish core collections. A group of scientists in China has constructed a core collection for Chinese soybean germplasm (Qiu Lijuan, personal com-

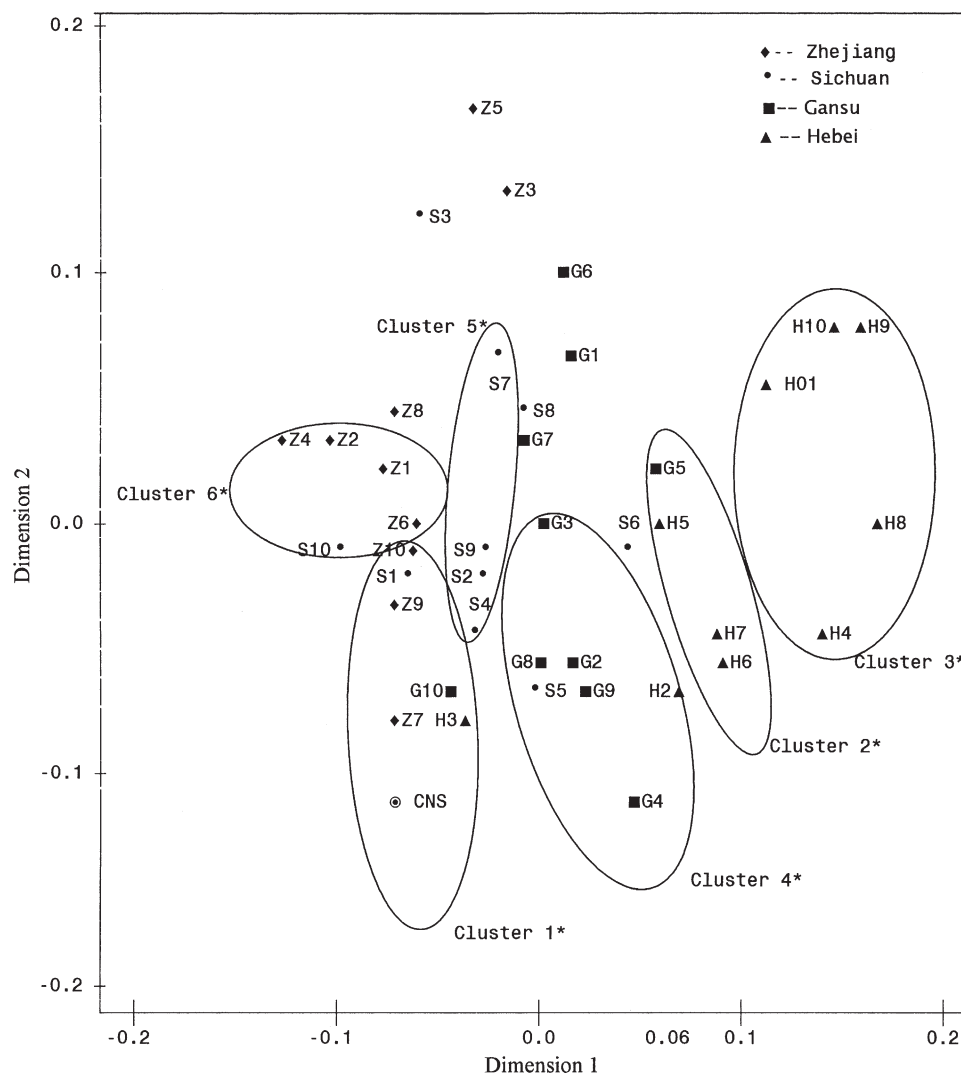


Fig. 2. Two-dimensional representation of genetic relations among 40 Chinese soybean germplasm accessions from Gansu (G), Hebei (H), Sichuan (S) and Zhejiang (Z) and the cultivar CNS derived from a multi-dimensional scaling (MDS) analysis. Genetic distance estimates are based on Jaccard's genetic dissimilarity matrix of 241 RAPD fragments generated by 31 primers. The circles were drawn to indicate the consensus clusters presented in Table 1.

munication, 2003). The sampling strategies for establishing their core collection were primarily based on the geographical information and morphological characters. The data obtained from our study have revealed ecogeographical distribution patterns of genetic variation that could be used to develop sampling strategies for establishing a core collection. For example, using the province of origin could be a useful criterion in developing a sampling strategy for accessions from China, but it would not be effective for accessions from Japan or South Korea where Li and Nelson (2001) found no relationship between genetic relationships, measured by RAPD fragments, and province of origin. This information would be particularly critical for the USDA Soybean Germplasm Collection, which has a nearly 20-fold difference between the provinces of China with the least and greatest number of accessions. Without a meaningful stratified sampling strategy, the important genetic diversity from the provinces with fewer accessions would be under-represented in the core collection. These data also show

that there are significant differences in the genetic distances when comparing accessions from pairs of provinces. We did not find a significant genetic difference between the accessions from Zhejiang and Sichuan, so it could be possible to treat the accessions from those two provinces as coming from a single source.

REFERENCES

- Aldrich, P.R., and J. Doebley. 1992. Restriction fragment variation in the nuclear and chloroplast genomes of cultivated and wild *Sorghum bicolor*. *Theor. Appl. Genet.* 85:293-302.
- Amadou, H.I., P.J. Bebeli, and P.J. Kaltsikes. 2001. Genetic diversity in Bambara groundnut (*Vigna subterranea* L.) germplasm revealed by RAPD markers. *Genome* 44:995-999.
- Anderson, J.A., G.A. Churchill, J.E. Autrique, S.D. Tanksley, and M.E. Sorrells. 1993. Optimizing parental selection for genetic linkage maps. *Genome* 36:181-186.
- Beebe, S., P.W. Skroch, J. Tohme, M.C. Duque, F. Pedraza, and J. Nienhuis. 2000. Structure of genetic diversity among common bean landraces of Middle American origin based on correspondence analysis of RAPD. *Crop Sci.* 40:264-273.
- Belaj, A., Z. Catovix, L. Rallo, and I. Trujillo. 2002. Genetic diversity

- and relationships in olive (*Olea europaea* L.) germplasm collections as determined by randomly amplified polymorphic DNA. *Theor. Appl. Genet.* 105:638–644.
- Bernard, R.L., C.R. Cremeens, R.L. Cooper, F.L. Collins, O.A. Krober, K.L. Athow, F.A. Laviolette, C.J. Coble, and R.L. Nelson. 1998. Evaluation of the USDA Soybean germplasm collection: Maturity groups 000 to IV (FC 01.547-PI 266.807). USDA Tech. Bull. 1844.
- Bhattacharjee, R., P. Bramel, C. Hash, M. Kolesnikova-Allen, and I. Khairwal. 2002. Assessment of genetic diversity within and between pearl millet landraces. *Theor. Appl. Genet.* 105:666–673.
- Bretting, P.K., and M.P. Widrechner. 1995. Genetic markers and plant genetic resource management. Vol. 13. p. 11–87. *In* J. Janick (ed.) *Plant breeding reviews*. John Wiley & Sons, New York.
- Brown, A.H.D., J.P. Grace, and S.S. Speer. 1987. Designation of a core collection of perennial *Glycine*. *Soybean Genet. Newsl.* 14:59–70.
- Brown, A.H.D. 1989a. The case of core collections. p. 135–156. *In* A.H.D. Brown et al. (ed.) *The use of plant genetic resources*. Cambridge Univ. Press, Cambridge, UK.
- Brown, A.H.D. 1989b. Core collections: A practical approach to genetic resources management. *Genome* 31:818–824.
- Brown-Guedira, G.L., J.A. Thompson, R.L. Nelson, and M.L. Warburton. 2000. Evaluation of genetic diversity of soybean introductions and North American ancestors using RAPD and SSR markers. *Crop Sci.* 40:815–823.
- Chang, R.Z., and J.Y. Sun. 1991. Chinese soybean collection catalog (Continued 1) (in Chinese). China Agricultural Press, Beijing, China.
- Chang, R.Z., and J.Y. Sun. L.J. Qiu, and Y. Chen. 1996. Chinese soybean collection catalog (Continued 2) (in Chinese). China Agricultural Press, Beijing, China.
- Ercan, A.G., M. Taskin, and K. Turgut. 2004. Analysis of genetic diversity in Turkish sesame (*Sesamum indicum* L.) populations using RAPD markers. *Genet. Res. Crop Evol.* 51:599–607.
- Excoffier, L., P.E. Smouse, and J.M. Quattro. 1992. Analysis of molecular variance inferred from metric distance among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics* 131:479–491.
- Fahima, T., G.L. Sun, A. Beharav, T. Krugman, A. Beiles, and E. Nevo. 1999. RAPD polymorphism of wild emmer wheat populations, *Triticum dicoccoides*, in Israel. *Theor. Appl. Genet.* 98:434–447.
- Fernandez, M.E., A.M. Figueiras, and C. Benito. 2002. The use of ISSR and RAPD markers for detecting DNA polymorphism, genotype identification and genetic diversity among barley cultivars with known origin. *Theor. Appl. Genet.* 104:845–851.
- Frankel, O.H. 1984. Genetic perspective of germplasm conservation. p. 161–170. *In* W. Arber et al. (ed.) *Genetic manipulation: Impact on man and society*. Cambridge University Press, Cambridge, UK.
- Gizlice, Z., T.E. Carter, Jr., and J.W. Burton. 1994. Genetic base for North American public soybean cultivars released between 1947 and 1988. *Crop Sci.* 34:1143–1151.
- Gizlice, Z., T.E. Carter, Jr., T.M. Gerig, and J.W. Burton. 1996. Genetic diversity patterns in North American public soybean cultivars based on coefficient of parentage. *Crop Sci.* 36:753–765.
- Iruela, M., J. Rubio, J.I. Cubero, J. Gil, and T. Millan. 2002. Phylogenetic analysis in the genus *Cicer* and cultivated chickpea using RAPD and ISSR markers. *Theor. Appl. Genet.* 104:643–651.
- Johnson, R.A., and D.W. Wichern. 1992. *Applied multivariate statistical analysis*. Prentice Hall, Englewood Cliffs, NJ.
- Keim, P., T.C. Olson, and R.C. Shoemaker. 1988. A rapid protocol for isolating soybean DNA. *Soybean Genet. Newsl.* 15:150–152.
- Kresovich, S., W.F. Lamboy, R. Li, J. Ren, A.K. Szewc-McFadden, and S.M. Blik. 1994. Application of molecular methods and statistical analyses for discrimination of accessions and clones of vetiver grass. *Crop Sci.* 34:805–809.
- Li, Z., L. Qiu, J.A. Thompson, M.M. Welsh, and R.L. Nelson. 2001. Molecular genetic analysis of U.S. and Chinese soybean ancestral lines. *Crop Sci.* 41:1330–1336.
- Li, Z., and R.L. Nelson. 2001. Genetic diversity among soybean accessions from three countries measured by RAPDs. *Crop Sci.* 41:1337–1347.
- Li, Z., and R.L. Nelson. 2002. RAPD marker diversity among soybean and wild soybean accessions from four Chinese provinces. *Crop Sci.* 42:1737–1744.
- Lopez-Valenzuela, J.A., O. Martinez, and O. Paredes-Lopez. 1997. Geographic differentiation and embryo type identification in *Mangifera indica* L. cultivars using RAPD markers. *HortScience* 32:1105–1108.
- Mellish, A., B. Coulman, and Y. Fernandez. 2002. Genetic relationships among selected crested wheatgrass cultivars and species determined on the basis of AFLP markers. *Crop Sci.* 42:1662–1668.
- Menkir, A., P. Goldsbrough, and G. Ejeta. 1997. RAPD based assessment of genetic diversity in cultivated races of sorghum. *Crop Sci.* 37:564–569.
- Nelson, R.L., P.J. Amdor, J.H. Orf, J.W. Lambert, J.F. Cavins, R. Kleiman, F.A. Laviolette, and K.A. Athow. 1987. Evaluation of the USDA soybean germplasm collection: Maturity groups 000 to IV (PI 273.483 to PI 427.107). USDA Tech. Bull. 1718.
- Nelson, R.L., P.J. Amdor, J.H. Orf, and J.F. Cavins. 1988. Evaluation of the USDA soybean germplasm collection: Maturity groups 000 to IV (PI 427.136 to PI 445.845). USDA Tech. Bull. 1726.
- Reif, J.C., A.E. Melchinger, X.C. Xia, M.L. Warburton, D.A. Hoisington, S.K. Vasal, G. Srinivasan, M. Bohn, and M. Frisch. 2003. Genetic distance based on simple sequence repeats and heterosis in tropical maize populations. *Crop Sci.* 43:1275–1282.
- SAS Institute. 1999. *SAS/STAT user's guide*, Version 8.0, First ed., SAS Inst., Inc., Cary, NC.
- Schneider, S., J.M. Kueffer, D. Roessli, and L. Excoffier. 2000. Arlequin ver. 2.000: A software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland.
- Schoen, D.J., and A.H.D. Brown. 1995. Maximising allelic diversity in core collections of wild relatives: The role of genetic markers. p. 55–76. *In* T. Hodgkin et al. (ed.) *Core collections of plant genetic resources*. John Wiley & Sons, Chichester, UK.
- Shepard, R.N. 1974. Representation of structure in similarity data: Problems and prospects. *Psychometrika* 39:373–421.
- Sneller, C.H. 1994. Pedigree analysis of elite soybean lines. *Crop Sci.* 34:1515–1522.
- Spagnoletti Zeuli, P.L., and C.O. Qualset. 1993. Evaluation of five strategies for obtaining a core subset from a large genetic resource collection of durum wheat. *Theor. Appl. Genet.* 87:295–304.
- Thompson, J.A., and R.L. Nelson. 1998. Core set of primers to evaluate genetic diversity in soybean. *Crop Sci.* 38:1356–1362.
- Thompson, J.A., R.L. Nelson, and L.O. Vodkin. 1998. Identification of diverse soybean germplasm using RAPD markers. *Crop Sci.* 38:1348–1355.
- Thormann, C.E., and T.C. Osborn. 1992. Use of RAPD and RFLP markers for germplasm evaluation. *In* *Application of RAPD technology to plant breeding*. Minneapolis, MN.
- vom Brocke, K., A. Christinck, E. Weltzien R., T. Presterl, and H.H. Geiger. 2003. Farmers' seed systems and management practices determine pearl millet genetic diversity patterns in semiarid regions of India. *Crop Sci.* 43:1680–1689.
- Wang, C.L. 1982a. Soybean (in Chinese). Heilongjiang Science and Technology Press, China.
- Wang, G.X. 1982b. Chinese soybean collection catalog (in Chinese). China Agricultural Press.
- Weir, B. 1990. *Genetic data analysis: Methods for discrete population genetic data*. Sinauer Assoc., Sunderland, MA.
- Yee, E., K.K. Kidwell, G.R. Sills, and T.A. Lumpkin. 1999. Diversity among selected *Vigna angularis* (Azuki) accessions on the basis of RAPD and AFLP markers. *Crop Sci.* 39:268–275.
- Zhang, D.P., J. Cervantes, Z. Huaman, E. Carey, and M. Ghislain. 2000. Assessing genetic diversity of sweet potato (*Ipomoea batatas* (L.) Lam.) cultivars from tropical America using AFLP. *Genet. Res. Crop Evol.* 47:659–665.